Endogeneity and instrumental variables

ECON306 – Slides 4
Studenmund Ch. 6 and 14

Bruno Salcedo

The Pennsylvania State University

Summer 2014

[0]

Motivation

- We consider linear models of the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- We compute estimates $\hat{\beta}_1$ to guide choices based on:

$$\Delta y_i = \hat{\beta}_1 \Delta x_i$$

- For this information to be useful it is important that $\hat{\beta}_1 \approx \beta_1$
- Under the classical assumptions we can guarantee:
  - Unbiasedness $\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$
  - Consistency $\hat{\beta}_1 \xrightarrow[\text{p}]{} \beta_1$

Endogeneity

- One of the crucial assumptions is orthogonality
- x_i is exogenous when it is uncorrelated with the error term:

$$\mathbb{E}[\, x_i \varepsilon_i \,] = 0$$

  Otherwise it is endogenous

- Stronger notions of exogeneity require:
  - ε to be independent from x
  - or conditional mean independence $\mathbb{E}[\, \varepsilon | x \,] = 0$
- Exogeneity guarantees that the things that are not accounted for (ε_i), do not interfere with the estimation of β_1

Why doe we care?

- Orthogonality is the most important and delicate assumption

- Failures of other assumptions can be tested (to some degree)
- Data from a model with a high degree of endogeneity can look completely normal
- Endogeneity can only be established/assumed through common sense/theory

- If other assumptions fail, we can still estimate β_1 consistently, and we can make inference with minor adjustments
- Endogeneity does not allow to estimate β_1 consistently
- If there is strong endogeneity bias, our estimated models can be poor descriptions of reality

Endogeneity bias

- Recall (see slides 3) that we can write:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \beta_1 + \frac{\sum(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum(x_i - \bar{x})^2}$$

- Which implies that:

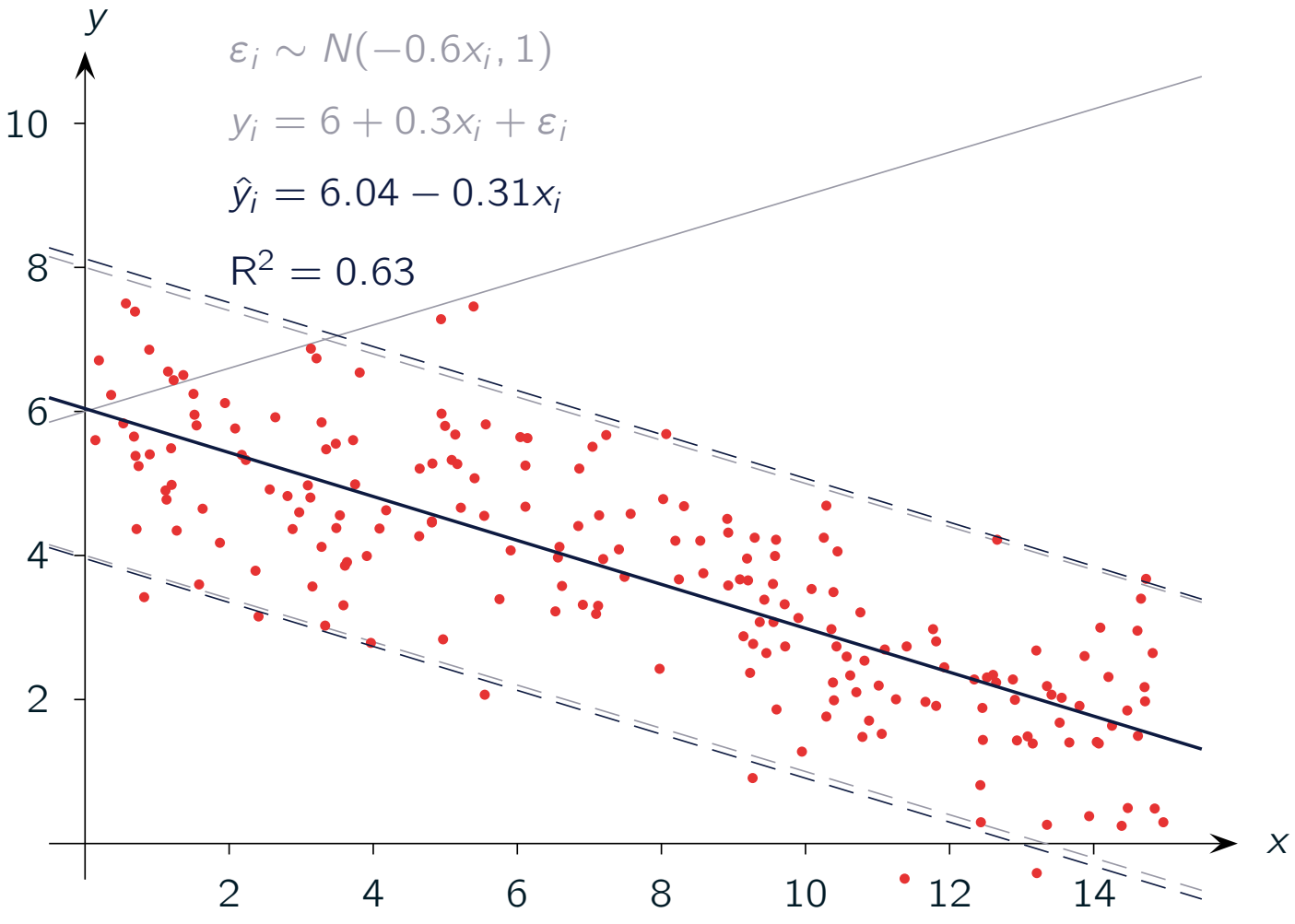
$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \mathbb{E}\left[\frac{\sum(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum(x_i - \bar{x})^2}\right] = 0$$

- And $\hat{\beta}_1$ is an unbiased estimator of β_1 only if:

$$\mathbb{E}\left[\frac{\sum(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum(x_i - \bar{x})^2}\right] = 0$$

Example: correlation between x and ε

Estimated model



Example: correlation between x and ε

The real problem

- We could define $\eta_i = \varepsilon_i + 0.6x_i$ and rewrite the model as:

$$\begin{aligned} y_i &= 6 + 0.3x_i + \varepsilon_i \\ &= 6 + 0.3x_i + (\eta_i - 0.6x_i) = 6 - 0.3x_i + \eta_i \end{aligned}$$

- Notice that $\eta_i \sim N(0, 1)$ and $\mathbb{E}[\, x_i \eta_i \,] = 0$
- All classical assumptions are satisfied!

- Why is endogeneity so important then?
  - x_i could be police officers, y_i could be crime rate and ε_i could be demographics
  - A policy that changes x_i may have no effect on ε_i

- Endogeneity is important when we can influence x_i but not ε_i

[0]

Omitted variables

- The most common source of endogeneity is omitted variables
- One of the reason for having an error term ε_i, is because we cannot account for all determinants of y_i
- For our estimates to be consistent, the effect of x_i must not depend on those things that we omit

- Suppose data comes from (with all classical assumptions satisfied):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \eta_i$$

- And instead we estimate:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

- We omit x_2 (often because we cannot observe it)

Omitted variable bias

- After some algebra:

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \beta_2 \mathbb{E}\left[\frac{\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum(x_{1i} - \bar{x}_1)^2}\right] + \mathbb{E}\left[\frac{\sum(x_{1i} - \bar{x}_1)(\eta_i - \bar{\eta})}{\sum(x_{1i} - \bar{x}_1)^2}\right] \to 0$$

$$= \beta_1 + \beta_2 \mathbb{E}\left[\frac{\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum(x_{1i} - \bar{x}_1)^2}\right]$$

- The term in the expectation is the OLS estimator of α_1 in the model:

$$x_{2i} = \alpha_0 + \alpha_1 x_{1i} + \varphi_i$$

- Therefore we have that:

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \beta_2 \mathbb{E}\left[\hat{\alpha}_1\right] = \beta_1 + \underbrace{\beta_2 \alpha_1}_{\text{bias}}$$

- The OLS estimator of β_1 will be consistent only if there is no bias, i.e.
  - If the omitted variable has no linear relation with y_i ($\beta_2 = 0$)
  - Or if the omitted variable has no linear relation with x_i ($\alpha_1 = 0$)

Example: the returns of schooling

Omitted variable

- Many people are interested in estimating the effect of additional years of schooling (EDU) on earnings (WAGE)
- Typically, people use models of the form

$$\text{WAGE}_i = \beta_0 + \beta_1 \, \text{EDU}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

  where $\mathbf{x}_i$ represents a vector of control variables (age, gender, parent's wealth, parent's education, ...)

- This usually involve some endogeneity because of unobserved characteristics such as innate talents or innate productivity (SKILL)

Example: the returns of schooling

Skill and education

- One would expect that more productive people have lower costs and expect higher returns from schooling
- Hence more productive people are prone to stay in school longer
- We would expect a positive relationship between EDU and SKILL, i.e. $\alpha_1 > 0$ in the model

$$\text{EDU}_i = \alpha_0 + \alpha_1 \, \text{SKILL}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

Example: the returns of schooling

Skill and earnings

- One would expect that more productive people have higher earnings controlling for other factors
- We would expect a positive relationship between WAGE and SKILL, i.e. $\beta_2 > 0$ in the model

$$\text{WAGE}_i = \beta_0 + \beta_1 \, \text{EDU}_i + \beta_2 \, \text{SKILL}_i + \boldsymbol{\gamma} \mathbf{x}_i + \varepsilon_i$$

Example: the returns of schooling

Omitted variable bias

- From the previous analysis, if we estimate the incomplete model

$$\text{WAGE}_i = \beta_0 + \beta_1 \, \text{EDU}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

- We should expect a positive bias:

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \alpha_1 \beta_2 > \beta_1$$

- That is, we would overestimate the effect of education
- This happens because it may be the case that people with more years of schooling would have higher wages, *even if they had't gone to school longer*, just because they are more productive
- Since innate skills are not affected by schooling, the direct of β_1 can result in poor policy recommendations

Example: smoking during pregnancy

Omitted variable

- It is now accepted that smoking during pregnancy (SMOKE) can result in low weight of the baby (WEIGHT)
- You can often find a warning in cigarette packages, but establishing this fact took many years of research

- One could simply use a model

$$\text{WEIGHT}_i = \beta_0 + \beta_1 \, \text{SMOKE}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

- There could be unobserved characteristics of the mother, that may affect WEIGHT and may be correlated with the decision to smoke
- We will consider the concern of the mother about her and her baby's health (HEALTH)

Example: smoking during pregnancy

Omitted variable bias

- Women with high values of HEALTH, are prone to not smoke, or to stop smoking when they learn they are pregnant.
- Hence we could expect a negative relation between HEALTH and SMOKE, i.e. $\alpha_1 < 0$ in:

$$\text{SMOKE}_i = \alpha_0 + \alpha_1 \, \text{HEALTH}_i + \boldsymbol{\gamma} \mathbf{x}_i + \varepsilon_i$$

- High valued of health might imply that the mother is healthy, and takes a number of measures to ensure the health of the baby
- Hence we may expect a positive relation between HEALTH and WEIGHT, i.e. $\beta_2 > 0$ in:

$$\text{WEIGHT}_i = \beta_0 + \beta_1 \, \text{SMOKE}_i + \beta_2 \, \text{HEALTH}_i + \boldsymbol{\gamma} \mathbf{x}_i + \varepsilon_i$$

Example: smoking during pregnancy

Omitted variable bias

- From the previous analysis, if we estimate:

$$\text{WEIGHT}_i = \beta_0 + \beta_1 \, \text{SMOKE}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

  without including HEALTH, we should expect a downward bias:

$$\mathbb{E}\left[\, \hat{\beta}_1 \,\right] = \beta_1 + \alpha_1 \beta_2 < \beta_1$$

- People who smoke tend to have other unhealthy habits
- It is hard to tell whether babies of smokers weight less because or smoking or because of something else
- This sort of concern may prevent the FDA from regulating the tobacco industry

Simultaneous equations

- Many Economic models are based on notions of equilibrium, that relate variables with more than one equation

- Suppose that data is generated according to:

$$y_{1i} = \beta_0 + \beta_1 y_{2i} + \gamma \mathbf{x}_i + \varepsilon_i$$
$$y_{2i} = \alpha_0 + \alpha_1 y_{1i} + \theta \mathbf{z}_i + \eta_i$$

- Notice that a change of ε_i leads to an increase Δy_{1i}, which in turns leads to an increase Δy_{2i}
- Hence ε_i and y_{2i} are correlated!

- This is where the distinction endogeneity vs. exogeneity comes from:
  - We think of $\mathbf{x}$, ε an η as exogenous variables (determined first)
  - And of y_{1i} and y_{2i} as endogenous variables (determined second)

Example: market equilibrium

endogenous variables

- The quantity demanded (D) and the quantity supplied (S) depend on the market price as well as on other factors:

$$\mathrm{D}_i = \beta_0 + \beta_1 \mathrm{P}_i + \gamma \mathbf{x}_i + \varepsilon_i$$
$$\mathrm{S}_i = \alpha_0 + \alpha_1 \mathrm{P}_i + \theta \mathbf{z}_i + \eta_i$$

- The price is determined in equilibrium as to clear the markets:

$$\mathrm{D}_i = \mathrm{S}_i$$

- Three equations and three "endogenous" variables D, S and P

Example: market equilibrium

elasticities

- Policy recommendations (eg. optimal prices or taxes) may depend on the elasticity of demand β_1 and the elasticity of supply α_1
- We could try yo estimate β_1 using OLS for the regression

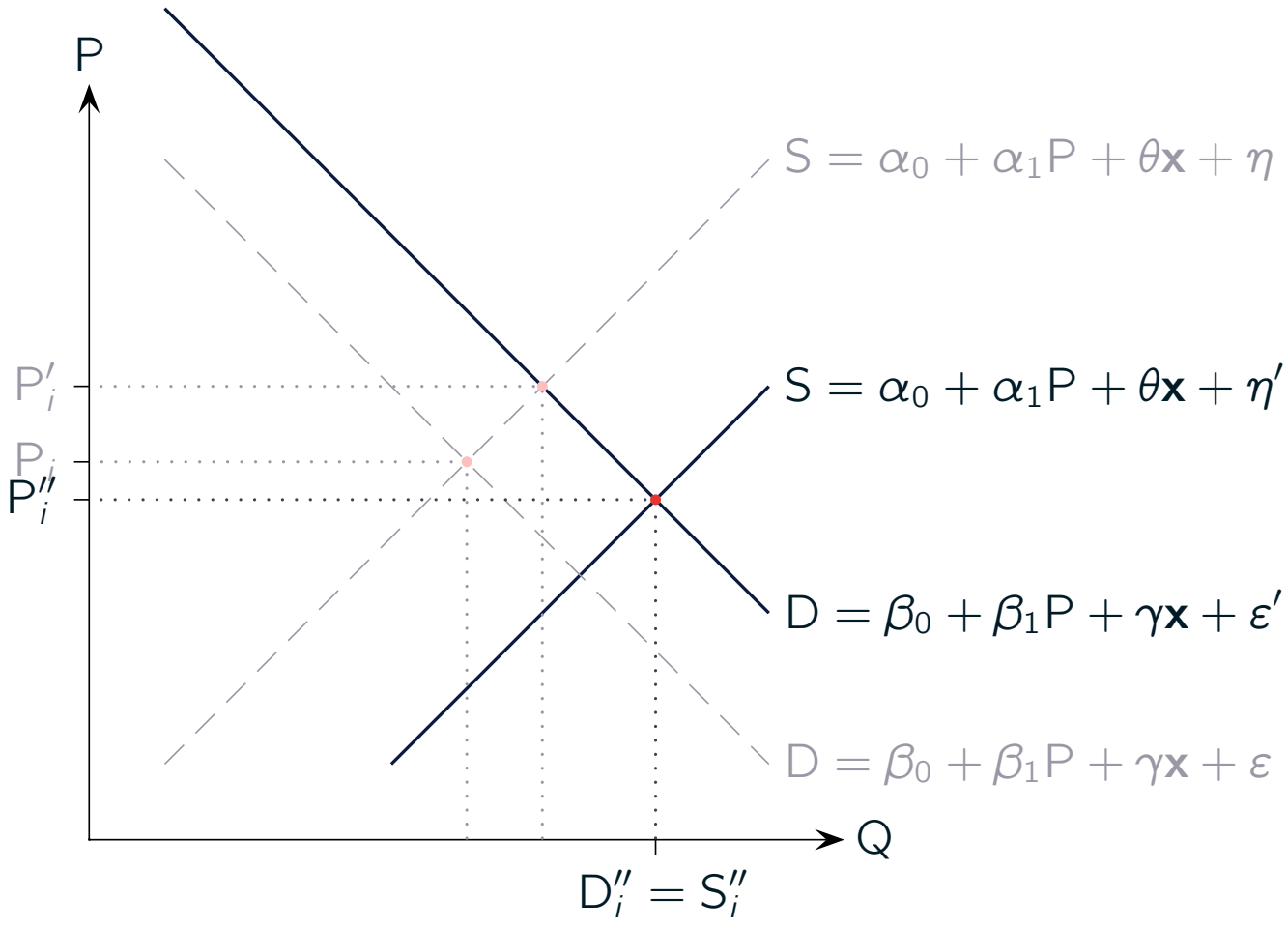
$$\mathrm{D}_i = \beta_0 + \beta_1 \mathrm{P}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

- But the law of demand ($\alpha_1 > 0$) suggests that $\mathbb{E}[\,\varepsilon_i \mathrm{P}_i\,] > 0$
- Which would result in biased estimates

$$\mathbb{E}\left[\,\hat{\beta}_1\,\right] = \beta_1 + \frac{\mathbb{C}[\,\mathrm{P}_i\,,\,\varepsilon_i\,]}{\mathbb{V}[\,\mathrm{P}_i\,]} > \beta_1$$

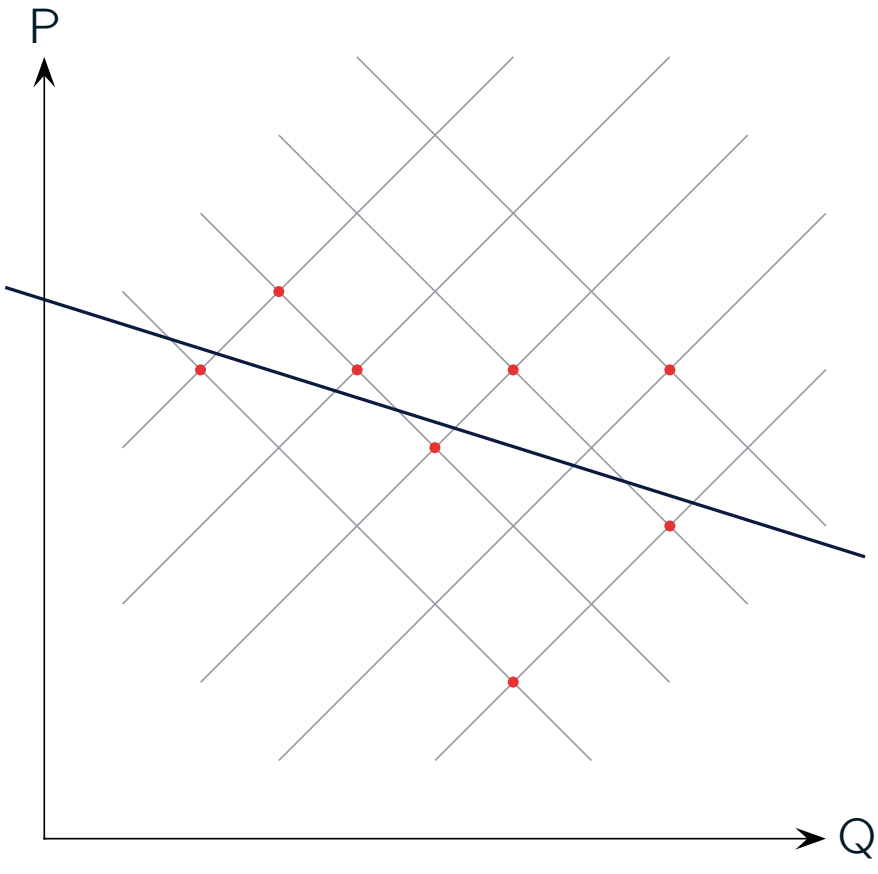
Example: market equilibrium

changes in η and ε lead to changes in P and Q



Example: market equilibrium

estimated model



$$\hat{Q} = 22.00 - 3.19\,P$$

Reduced form

- One form of dealing with simultaneous equations is to identify which variables are endogenous
- Then one can simply include the exogenous variables from all equations as regressors

- Instead of estimating:

$$y_{1i} = \beta_0 + \beta_1 y_{2i} + \gamma \mathbf{x}_i + \varepsilon_i$$
$$y_{2i} = \alpha_0 + \alpha_1 y_{2i} + \theta \mathbf{z}_i + \eta_i$$

- One could estimate the reduced form model:

$$y_{1i} = \beta_0 + \gamma \mathbf{x}_i + \lambda \mathbf{z}_i + \varepsilon_i$$
$$y_{2i} = \beta_0 + \theta \mathbf{x}_i + \theta \mathbf{z}_i + \eta_i$$

- With this approach is we cannot estimate β_1 nor α_1!

Example: demand and supply revisited

- We have three endogenous variables S, D and P
- We could simply estimate the reduced form equations:

$$\mathrm{S}_i = \mathrm{D}_i = \beta_0 + \gamma \mathbf{x}_i + \lambda \mathbf{z}_i + \varepsilon_i$$
$$\mathrm{P}_i = \pi_0 + \theta \mathbf{x}_i + \theta \mathbf{z}_i + \eta_i$$

- This could tell us how different exogenous variables in $\mathbf{x}$ and $\mathbf{z}$ affect the market outcome

- But we cannot recover the most important structural parameters: β_1 and α_1 which describe the demand and supply curves

Example: police surveillance

- Police presence/surveillance (COPS) can potentially reduce crime rates (CRIME)
- One could measure this effect with a simple model:

$$\text{CRIME}_i = \beta_0 + \beta_1 \, \text{COPS}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

- However COPS may be endogenous since police departments distribute target their resources to address problematic areas
- One may expect a significant coefficient α_1 in:

$$\text{COPS}_i = \alpha_0 + \alpha_1 \, \text{CRIME}_i + \theta \mathbf{z}_i + \eta_i$$

- Simultaneity makes it hard to identify β_1

Example: welfare and political stability

- Civil wars are still common in present days.
- Understainding which factors affect political stability, may help to prevent them.

- One may think that an important factor in predicting conflicts (WAR), is the state of the economy (GDP).
- In prosperous times, people are less likely to be discontent or antagonize each other.

- Estimating the effect of GDP on WAR may be hard, because there may be backward causality
    - Civil wars have devastating effects for the economy
    - Instability makes a country less attractive for investors
- Hence we can expect that WAR also affects GDP

Selection bias

- Suppose we want to estimate the effect of a treatment denoted by a dummy variable d:

$$y_i = \beta_0 + \beta_1 d_i + \varepsilon_i$$

- Notice that:

$$\mathbb{E}[\, y_i | d_i = 1 \,] = \beta_0 + \beta_1 + \mathbb{E}[\, \varepsilon_i | d_i = 1 \,]$$
$$\mathbb{E}[\, y_i | d_i = 0 \,] = \beta_0 + \mathbb{E}[\, \varepsilon_i | d_i = 0 \,]$$

- Therefore

$$\begin{aligned} \mathbb{E}\left[\, \hat{\beta}_1 \,\right] &= \mathbb{E}[\, y_i | d_i = 1 \,] - \mathbb{E}[\, y_i | d_i = 0 \,] \\ &= \beta_1 + \Big(\mathbb{E}[\, \varepsilon_i | d_i = 1 \,] - \mathbb{E}[\, \varepsilon_i | d_i = 0 \,] \Big) \end{aligned}$$

- We obtain consistent estimates only if $\mathbb{E}[\, \varepsilon_i | d_i = 1 \,] = \mathbb{E}[\, \varepsilon_i | d_i = 0 \,]$
- That is, if the selection criteria for the treatment is orthogonal to ε

Example: the returns of military service

- It is important to understand what are the long-term consequences of military service (MS) for people who will not pursue military careers
- In particular one could be concerned about the effect on income (WAGE)

$$\text{WAGE}_i = \beta_0 + \beta_1 \, \text{MS}_i + \beta_2 \, \text{MS}_i \cdot \text{AGE}_i + \beta_3 \, \text{MS}_i \cdot \text{AGE}_i^2 + \boldsymbol{\gamma} \mathbf{x}_i + \varepsilon_i$$

- A potential problem is that the choice to participate in voluntary service is endogenous
- Candidates with less attractive outside options have lower opportunity cost and are more likely to participate
- This same candidates are likely to have higher wages *caeteris paribus*
- Hence it is likely that:

$$\mathbb{E}[\, \varepsilon_i | \text{MS}_i = 1 \,] \neq \mathbb{E}[\, \varepsilon_i | \text{MS}_i = 0 \,]$$

Example: program evaluation

- It is important to assess the effectiveness of different governmental program
- To do this, one could try to compare outcomes of regions in which the program is implemented, with the outcomes of regions where it is not
- A potential problem is that the choice of where to implement the program is endogenous
- Governments are likely to implement programs (first?) in regions where they are expected to be more effective
- The selection criteria may thus be correlated with the residuals

Measurement error

- The measurement of economic activity may be susceptible of measurement error
- Consider the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Suppose that instead of observing x, we observe $x^* = x + \mu$
- Hence the model that we actually estimate is:

$$y_i = \beta_0 + \beta_i x_i^* + \eta_i$$
$$\eta_i = \varepsilon_i - \beta_1 \mu_i$$

- In many cases, it is sensible to assume that the measurement error is correlated with the measured variable

Example: health and wealth

- Wealthier people tend to be healthier
- Better nutrition, better medial services ...

- Wealth data is often self reported
- Self-reports of wealth usually carry some measurement error
    - Most people don't know the exact value of their assets and income
    - Some people may want to distort information for privacy concerns

- Both effects tend to be higher for richer people
- Hence we may have $\mathbb{E}[\, x_i \mu_i \,] < 0$

[0]

The problem

- We wish to estimate the effect of x on y, namely β_1
- The problem with endogeneity is that:

  We cannot disentangle the variation of x from variation of other factors (ε) which may also affect y

- For example:
  - If we only observe prices and quantities, we cannot tell whether increases in quantities are due to increases in prices or to increases in demand
  - If we only observe earnings and years of schooling, we do not know if high earnings are due to education or skill

Natural experiments

- If we could observe variations of x which are independent from ε, we could then estimate β_1

- This is exactly what happens in controlled experiments in which the researcher varies x keeping everything else constant
- Controlled experiments are rarely possible in realistic economic scenarios
- Laboratory experiments in Economics are often questioned for lack of external validity

- An alternative is to look for natural experiments which affect x but nothing else
- The idea of natural experiments is materialized in instrumental variables

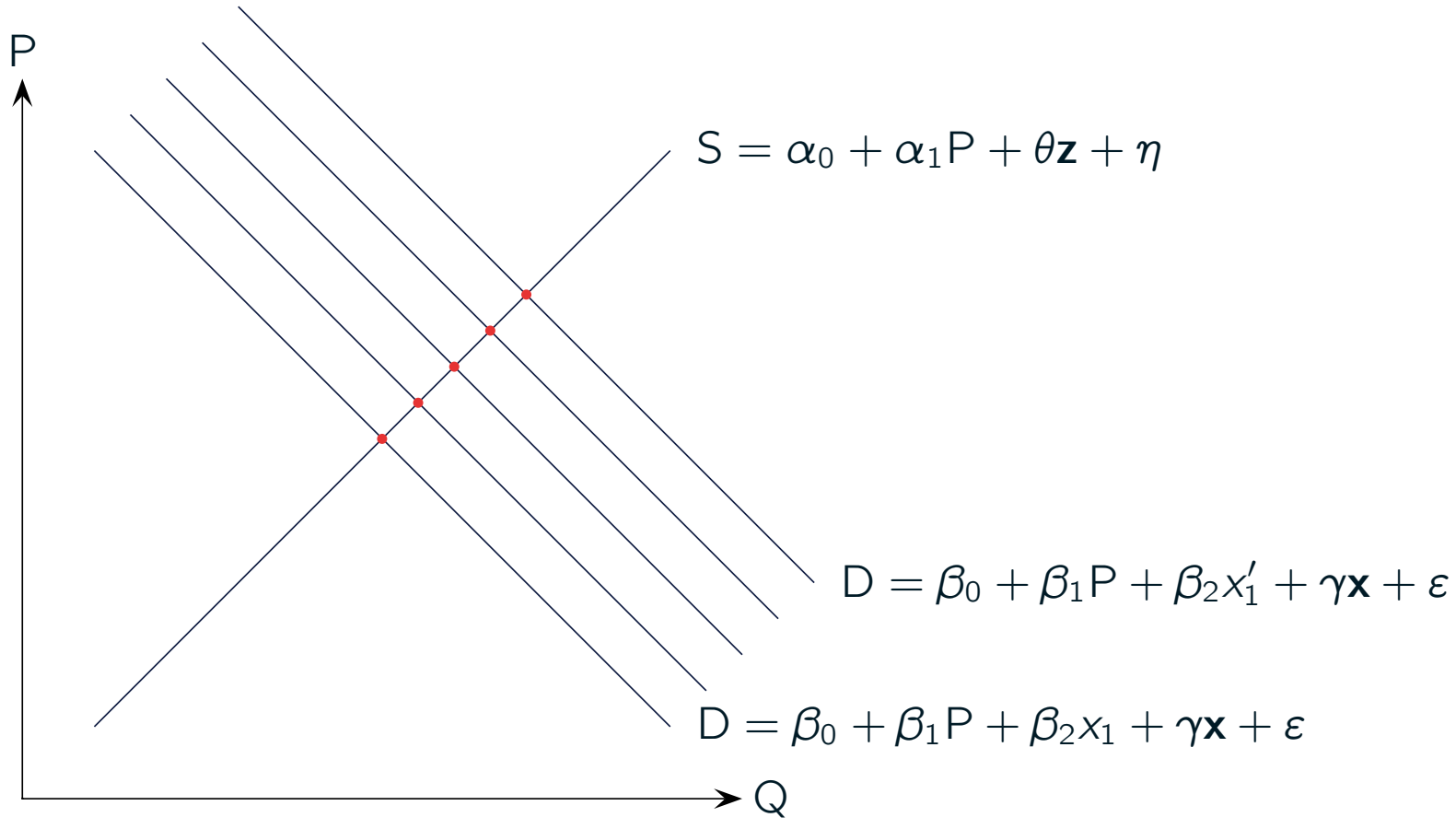
Instrumental variables

- An instrumental variable (IV) is a variable that is related to x but is otherwise independent from y
  - It affects y but only through x
  - In particular it is independent (orthogonal) to ε
- An example of an intrumental variable are the manipulations of a researcher in an experimental setting
- Instead of considering all the variation of x, focus on the variation explained by z
- This variation is exogenous!
- If this variation is large enough, we can use it to identify β_1
- In that sense IVs can serve as instruments to identify the effect of x on y

Example: market equilibrium

demand shifters

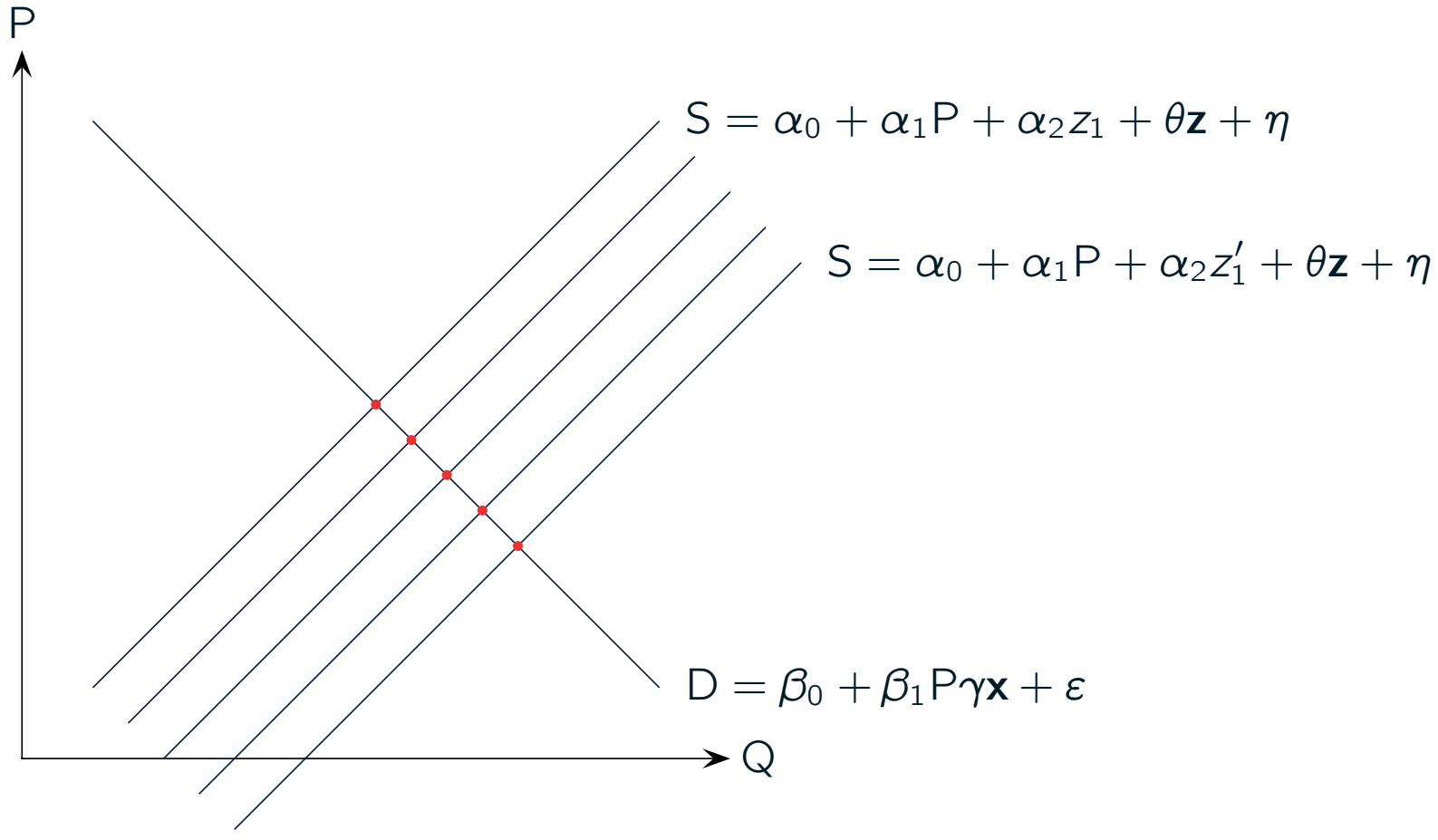
Suppose a variable x_1 affects demand but not supply
Then we can observe changes in P which are independent from η



Example: market equilibrium

supply shifters

Suppose a variable z_1 affects supply but not demand
Then we can observe changes in P which are independent from ε



Example: market equilibrium

demand and supply shifters

Demand	Supply
Population	Number of competitors
Income	Technological shocks
Preferences (ads,fads)	Prices of inputs
Expectations	Expectations
Prices of substitutes/complements	Prices of alternative goods

Two stage ordinary least squares

Setup

- Suppose that you want to estimate:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- However you are afraid that x may be endogenous

- To fix this problem, you use z as an instrument
- z should be related to x and independent from ε

- Instead of considering all the variation in x we can focus by the variation of x which can be explained by z
- We do this using two stage ordinary least squares (2SOLS)

Two stage ordinary least squares

First stage

- To capture the variation of x which can be explained by z, consider the model

$$x_i = \alpha_0 + \alpha_1 z_i + \eta_i$$

- The first step of 2SOLS is to run regular OLS on this model to obtain estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$

- We then use these estimates to compute predicted values:

$$\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i$$

which are independent from ε_i!!

Two stage ordinary least squares

Second stage

- The predictions $\hat{x}_i$ contain all the variation of x which can be explained by z
- Assuming that η and z are exogenous, then so is $\hat{x}$

- We can use $\hat{x}$ to estimate the effect of x over y by running OLS on

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \varphi_i$$

- The estimators obtained in the second stage are the IV estimators of β_0 and β_1
- They are often denoted by $\hat{\beta}_0^{\text{IV}}$ and $\hat{\beta}_1^{\text{IV}}$

Example: market equilibrium

2SOLS for demand

- Suppose that you want to estimate demand
- You need a supply shifter z_1 (e.g. price of inputs)

- On the first stage you would use OLS to estimate:

$$\mathrm{P}_i = \delta_0 + \delta_1 z_{1i} + \varphi_i$$

- Using your estimates you would construct predicted values:

$$\widehat{\mathrm{P}}_i = \hat{\delta}_0 + \hat{\delta}_1 z_{1i}$$

- These predictions capture variations in price which are independent from changes in demand

- The IV estimates for the demand function will be the OLS estimates for the model:

$$\mathrm{D}_i = \beta_0 + \beta_1 \widehat{\mathrm{P}}_i + \gamma \mathbf{x}_i + \varepsilon_i$$

Valid instruments

- Which variables can serve as instruments?
- A first obvious requirement is that it must be something which can be measured (data should be available)
- A good instrument should explain some of the variation of x
- The slope coefficient in the first stage model should be significant
- The R^2 coefficient should be high enough!
- A good instrument should be exogenous (orthogonal to ε)
- Unfortunately, endogeneity cannot be tested in general (tomorrow we will discuss an exception)
- The endogeneity of the instruments has to be established via common sense/theory

Examples of instrumental variables

Response	Regressor	Eondogeneity	Instruments
earnings	schooling	omitted	quarter of birth school construction proximity to college
newborn weight	smoking	omitted	random assignement state taxes
demand supply	price price	simultaneity simultaneity	supply shifters demand shifters
crime	police surveillance	simultaneity	electoral cycles Papal trajectory
conflict	GDP	simultaneity	rainfall
labor supply	fertility	simultaneity	gender composition of children
earnings	military service	selection	draft lottery

[0]